テキストマイニングのための統合環境「TETDM」のチュートリアル

砂山　渡
広島市立大学

本日の内容

17:00-17:05
– TETDMの紹介

17:05-17:40
– TETDMの利用体験

17:40-18:00
– TETDMの研究応用
– TETDMの教育応用

質問は随時受け付けます

TETDMプロジェクト

- 複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築
- 電子テキストを扱う多くのユーザの創造的活動を支援するツールの提供

2010年度から5年以内に達成する課題として
人工知能学会全国大会「近未来チャレンジ」の
プロジェクトとして発足，2015年度の大会で卒業認定！

テトリーヌ by えむたこ

テキストマイニングとは？

- テキストを利用した，雑多な目的に対する処理
  – 論文化され，有効性が検証された技術
  – 単語の頻度計算など基礎的な処理
  – 年齢や性別，出身地の推定など面白い処理
  – テキストが何らかの形で関わるデータマイニング
  – テキストが何らかの形で関わる処理

TETDMの活用場面

[利用者（プログラミングをしない人）]

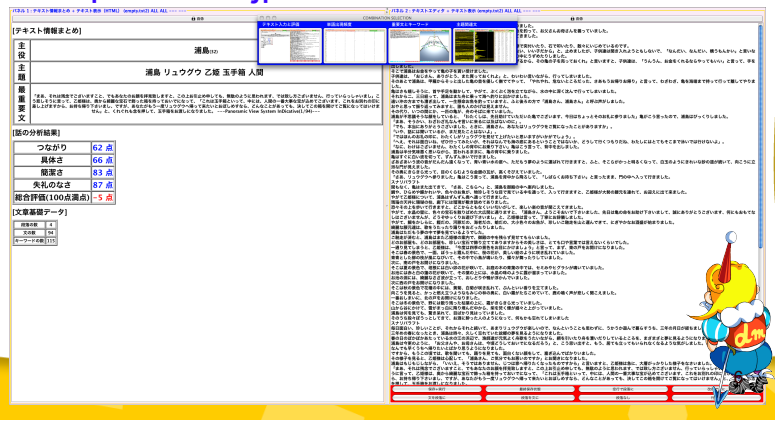
- 卒論，学会原稿，レポートなどの文章作成支援
- レポート，アンケート結果の分析支援
- メール，SNS，電子掲示板の文章のまとめや分析

[開発者（プログラミングする人）]

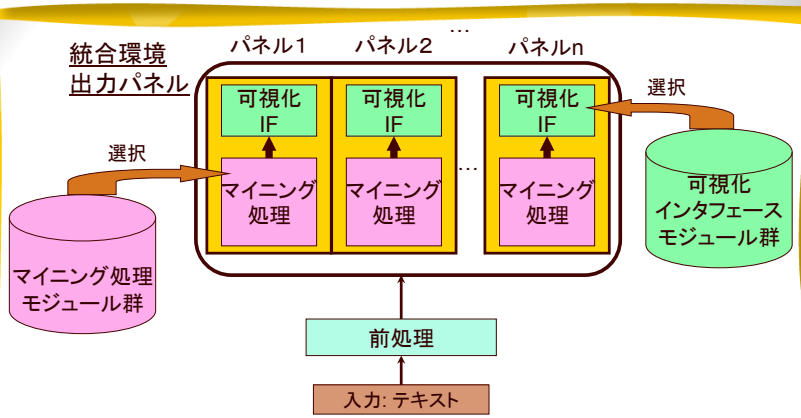
- 研究のためのシステム構築
- 授業や研究室における javaプログラミングや マイニングアルゴリズムの演習

TETDM(Ver.1.01)

- http://tetdm.jp/ よりダウンロード可能



TETDMの構成



TETDMの利用の流れ

1. TETDMの起動とテキスト入力
2. 起動モードを選択（本日は通常モードを利用）
3. パネルにツール（処理＋可視化）をセット
4. ツールを操作して分析する
5. 3.と 4.の繰り返しの中で，気になる結果を解釈する
6. 集めた解釈をもとに創発する

途中で時間に余裕のある方は，画面上部の チュートリアル「利用」ボタンで表示される チュートリアルもお試しください

1.TETDMの起動とテキスト入力

1) TETDM1024.jar をダブルクリック

2) TETDM起動後，右側のパネル（テキスト表示）に
文章をカット&ペーストで貼付ける

3) 「保存＋実行」ボタンを押す

4) 「空行で段落に」ボタンなどで，段落を生成する

TETDMの入力テキストの形式

- テキスト：日本語で書かれた文の集合

- 文の終わりに句点（全角）「．」「。」がある*
  *「キーワード設定」で指定の文字に変更可能
- 段落の終わりに文字列「スナリバラフト」がある*
  *「キーワード設定」で指定する任意の文字列に変更可能
- 日本語文字コードは，SHIFT-JISかEUC*
  *UTF-8は使えない．文字コードの自動判別は難しくjavaの問題

2.起動モードを選択

– 画面上部の「モード」ボタンを押して，モードを選択する
  モードごとに，利用できる機能やツールが異なる

- スーパーライトモード
  – 説明なしでの利用を想定（趣味での利用）
- ライトモード
  – 簡単な説明のみでの利用を想定（趣味での利用）
- 通常モード
  – 汎用性がある実用的な利用を想定（学業，仕事での利用）
- 拡張モード
  – 専門的な内容を含む利用を想定（学業，仕事での利用）
  – 独自のツールを開発したい人を想定

3.パネルにツールをセット

– 画面上部の「ツール自動組合せ」のためのウインドウから
用途に応じたボタンを押す
  - 単語出現頻度
  - 重要文とキーワード
  - 主題関連文
  - テキスト評価
  - 主語のない文
  …
  - 2ちゃんまとめ
  - タイピング（数字）

– パネル内部の「セット」ボタンを押す，「戻る」ボタンを押す
  - 各ツールの作成者が組み合わせて利用することを想定している
ツールを自動的にセット

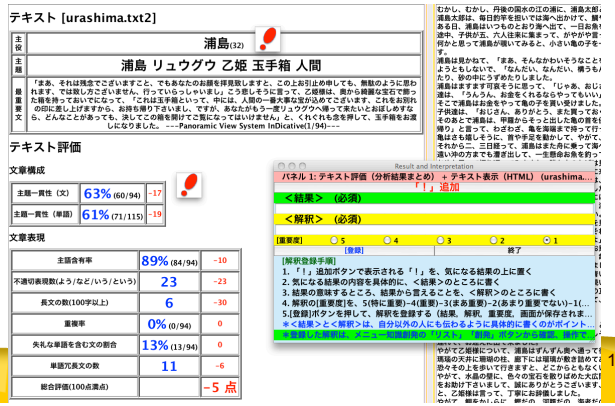
4.ツールを操作して分析する

- 操作の必要がない場合は，出力をそのまま見る
  - テキスト評価（分析結果まとめ）:「テキスト評価」
  - 長文抽出:「長文」
  - 失礼単語抽出:「失礼単語確認」
- 処理ツールは，主にボタンで操作する
  - テキストエディタ:「テキスト評価」
  - 文章要約（展望台）:「重要文とキーワード」
  - 主語抽出:「主語のない文」
- 可視化ツールは，主にマウスで表示内容を操作する
  - 表形式表示:「単語出現頻度」
  - キーワード表示（展望台）:「重要文とキーワード」（ツールの連動あり）
  - キーワード選択（フォーカス指定）:「主題関連文」（ツールの連動あり）

「」内はツール自動組合せの名称

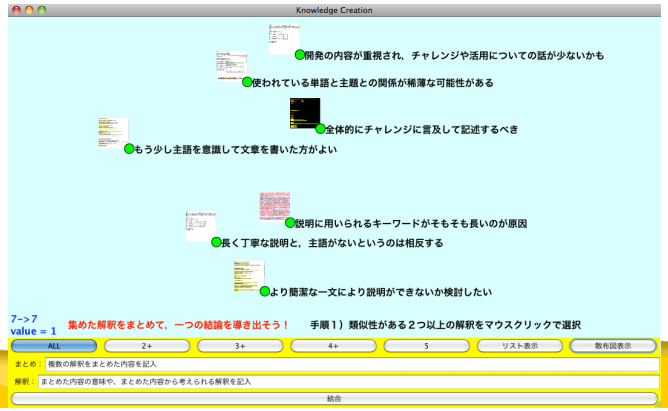
5.気になる結果を解釈する

- パネル内の「結果と解釈」ボタンを押して気になった結果とその解釈を登録する



6.集めた解釈をもとに創発する

- 画面上部の「創発」ボタンを押して表示されるインタフェース内で集めた解釈を1つにまとめる

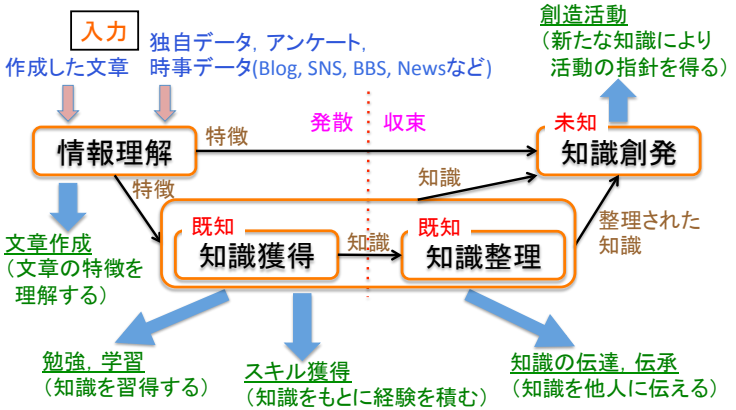


TETDMの研究応用

- 研究のツール作成
  - 研究の核となる部分のみの実装で動作
  - 形態素解析などの前処理の実装が不要
  - ウインドウ生成などの環境の実装が不要
  - 既存ツールの利用により，既存技術の利用や改変が容易
- 作成したツールの実用化
  - 作成したツールの提供が容易
  - 一般公開による利用者拡大の可能性

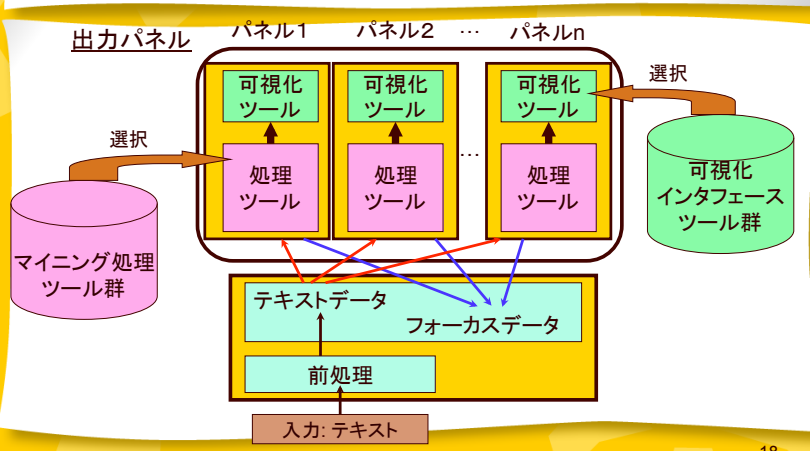
論文を投稿したら終わり！ではなく
実際に利用，再利用される機会を増やせる

テキストマイニング研究の主目的



情報：もの，こと，人，についての知識を含む可能性があるテキスト
知識：情報から導かれる事実（事象）やルール（事象間の関係）

TETDMの構成



TETDMの前処理

- 形態素解析：文章の単語への切り分け
  - 形態素解析器Igo: 解析結果はほぼMecab互換
- 単語の出現位置,頻度の取得
  - 文章を、文と段落に切り分け、
    各単語の出現する文、段落、とその頻度を計算
- 単語間、段落間の関連度の取得
  - 同じ段落に出現する単語情報から関連度
    （cos類似度）を計算

前処理後のテキストデータ

- 入力テキストの原文（段落，文ごとに区切る）
- 出現単語リスト（段落，文ごと）
- 単語種類数（段落，文ごと）
- 単語の出現頻度（総頻度，段落頻度，文頻度）
- 各単語の出現情報（出現する段落，文）
- インタフェース上で
  ユーザが注目している情報（段落，文，単語）(フォーカスデータ)

複数のテキストを入力した場合
1つのテキストを段落として扱う

処理ツールの作成方法

1. 組み合わせる可視化ツールの
   データ受け取り方法の確認
2. 処理の記述
3. 処理結果の送信部分の記述

例）名詞の頻度上位10個の単語を表示させたい

1.組み合わせる可視化ツールの データ受け取り方法の確認

可視化ツール「TextDisplay」

//データを受け取るメソッド

```
public boolean setData(int dataID, String t) //String型で受け取り
{
    switch(dataID)
    {
        case 0:    displayText = t;    //そのまま表示
                   return true;
    }
        return false;
}
```

String型でデータを送れば良い！

2.処理の記述

```
String MyMethod()     //独自の処理をさせるメソッドを作成
{
        int nounID[] = new int[text.keywordNumber];
        int frequency[] = new int[text.keywordNumber];
        int count;

        count = 0;
        for(int i=0;i<text.keywordNumber;i++)  //すべての単語
                if(text.keyword[i].partOfSpeech == 1)    //名詞
                {
                        frequency[i] = text.keyword[i].frequency;
                        count++;                          //頻度を配列に保存
                }
                else
                        frequency[i] = 0;

        Qsort.initializeIndex(nounID, text.keywordNumber);
        Qsort.quicksort(frequency, nounID, text.keywordNumber);
                                                          //頻度順にソート
```


3.処理結果の送信部分の記述

```
        StringWriter sw = new StringWriter();
        BufferedWriter bw = new BufferedWriter(sw);
        try{
                for(int i=0;i<10 && i<count;i++)
                        bw.write(text.keyword[nounID[i]].word+" ");
                                //頻度上位の名詞をスペース区切りで結合
                bw.flush();
        }
        catch(Exception e){
                System.out.println("writing ERROR in NounTop10");
        }
        return sw.toString();
}                                //バッファを利用しないで，より簡潔に書くことも可能
```

```
setDataString(MyMethod());
```

//String型のデータを送信

TETDMの教育応用

- 文章（卒論，修論，学会原稿）作成指導
  - 現状のTETDM（ツール「テキスト評価」など）をそのまま利用
  - 現状のTETDMをもとに，テキスト評価ツールを独自に作成
  - 自作のツールを加えて，テキスト評価ツールを独自に作成
- javaプログラミング演習
  - 雛形ツール（既存，独自）の改変によるツール作成
- テキストマイニングアルゴリズム演習
  - 必要なアルゴリズムのみの実装によるツール作成

少ない労力で実践的な学習，教育環境を構築できる

今後の展望

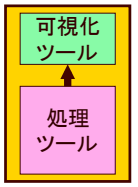
- 実社会応用の拡大
- 利用者／開発者向けコンテストの実施
- データマイニング用TETDM
- 英語版TETDM
- チュートリアルの拡張
- クラウドサーバを用いた協調的マイニング
- TETDMのゲーム化

TETDMプロジェクトメンバー

- 砂山渡(広島市立大学　大学院情報科学研究科)
- 高間康史(首都大学東京　システムデザイン学部)
- 西原陽子(立命館大学　情報理工学部)
- 徳永秀和(香川高等専門学校)
- 串間宗夫(宮崎大学　医学部附属病院医療情報部)
- 阿部秀尚(文教大学　情報学部)
- 梶並知記(神奈川工科大学　情報学部)
- 松下光範(関西大学　総合情報学部)
- Danushka Bollegala(ダヌシカ ボレガラ)(リバプール大学)
- 佐賀亮介(大阪府立大学　大学院工学研究科)
- 河原吉伸(大阪大学　産業科学研究所)
- 川本佳代(広島市立大学　大学院情報科学研究科)

ツールの種類

- マイニング処理ツール
  - 入力として与えられるテキストに関連して行われる処理全般
- 可視化インタフェースツール
  「汎用性を重視して可視化処理のみを実装」
  - マイニング処理の結果を視覚的に出力
  - マイニング処理や可視化の観点を変更するなど利用者が対話的な操作を行う入力インタフェース



パネル

ツールの例

- マイニング処理ツール
  - キーワード抽出
  - 文章要約
  - テキスト分類
  - 一貫性評価
- 可視化インタフェースツール
  - テキスト表示
  - 表形式表示
  - グラフ／ネットワーク表示
  - マップ表示

処理ツールと
可視化ツールを
ペアとして
パネルにセット

TETDMの特徴

[1.幅広い利用者と開発者の参入]
- 卑近なデータを入力可能，容易にツールの追加と利用が可能
  - 面白い，斬新な，任意のツールを追加，使用できる

[2.モジュール間での相互インタラクションの実現]
- 独立に作成された複数のモジュールを並列に並べられ
  ユーザの操作に対して協調動作，表示が可能
  - 既存研究は1つのグループが作成するシステムの中で協調表示

[3.知識創発のための基盤環境の構築]
- 処理＋可視化に加えて「解釈」「創発」を含めて支援
  - 既存研究は処理結果を提示するところまでに主眼がおかれている

文章の分析（全体構成）

- 文章の主題と一貫性
  - 処理「テキスト評価」＋可視化「テキスト表示(HTML)」→「セット」
  - 処理「文章要約（展望台）」＋可視化「テキスト表示」
  - 処理「主題関連文評価（光と影）」＋可視化「テキスト表示(カラー)」
  - 処理「主題関連語評価（川下り）」＋可視化「主題関連語表示（川形式）」
- 段落間の関係と構成
  - 処理「なし」＋可視化「段落間類似度表示」
  - 処理「なし」＋可視化「段落間木構造（類似度）」
  - 処理「なし」＋可視化「段落間木構造（トップダウン）」
  - 処理「段落順序評価（トップダウン）」＋可視化「段落並べ替え」

文章の分析（表現）

- 主語，曖昧表現，長文の有無
  - 処理「テキスト評価」＋可視化「テキスト表示(HTML)」→「セット」
  - 処理「主語抽出」＋可視化「テキスト表示(カラー)」
  - 処理「単語抽出(文章評価用)」＋可視化「テキスト表示(HTML)」
  - 処理「長文抽出」＋可視化「テキスト表示(HTML)」
  - 処理「類似文抽出」＋可視化「テキスト表示(カラー)」(100文まで)
- 使用単語の確認
  - 処理「なし」＋可視化「表形式表示」
  - 処理「なし」＋可視化「段落間ネットワーク(ばねモデル)」